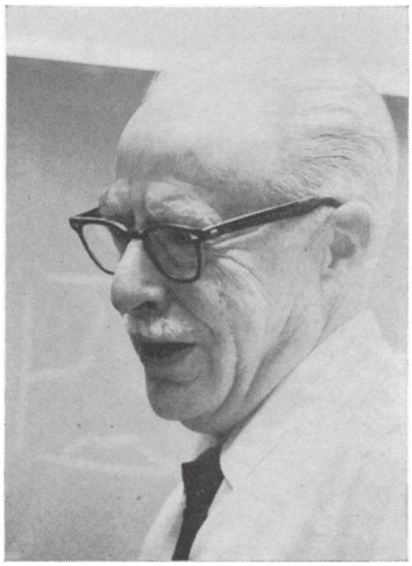
Genesis of the Transistor

Walter H. Brattain
Whitman College, Walla Walla, Washington 99362

As an experimental physicist, I have always been interested in exploring the frontiers of science to obtain new information that will contribute to a better understanding of nature. Frequently new inventions must await expansion of our fundamental knowledge before they can become a reality. The research that led to the discovery of the transistor effect and the invention of the transistor at Bell Telephone Laboratories in 1947 is a good example.

Walter H. Brattain was born in Amoy, China, on 10 February 1902. Educated in the State of Washington and receiving a B.S. degree from Whitman College in 1924, he was awarded the M.A. degree by the University of Oregon in 1926, and the Ph.D. by the University of Minnesota in 1929. Dr. Brattain was a member of the Bell Laboratories staff from 1929 to 1967. Upon his retirement in February 1967, he became visiting Professor of Physics at Whitman College in Walla Walla, Washington. The chief field of his research has been the surface properties of solids. His major contribution to solid state physics was his discovery of the photoeffect at the free surface of a semiconductor, and the invention of the point-contact transistor jointly with John Bardeen. Dr. Brattain is a member of the National Academy of Sciences and many scientific societies. He shared the Nobel Prize in Physics for 1956 with Dr. Bardeen and William Shockley for their investigations of semiconductors and the discovery of the transistor effect.



I would like to tell you the story as I remember it, for it may answer the question so often asked by science students, "How was the transistor invented?"

Before I begin, however, you may wish to know what a transistor is. The transistor's basic function is to regulate flow of current in response to an input signal. Briefly, it performs in a solid what the electron tube does in a vacuum: it conducts, modulates, switches, and amplifies electrical signals.

The transistor, like the vacuum tube, is an active circuit element. In a transistor we have a semiconductor in which both negative and positive charge carriers (electrons and holes, respectively) exist. The ability of the semiconductor solids to conduct electricity depends on the number of these holes and electrons.

Power from a local power source can be used to make a steady current flow through such a solid. If, for example, the electrical potential at one junction of contact is such that the flow of current is limited by the number of holes available and a potential is applied at a neighboring junction to increase the number of holes, this will increase the current flow. This process enables the transistor to amplify a weak, intelligible signal with very little expenditure of power.

Semiconductors such as germanium and silicon are ideal as active circuit elements because the number of electrons or holes they contain is very small. A few added holes (positive carriers) or added electrons will make a significant difference. Actually, added holes result also in added electrons or vice versa, since a conductor cannot become charged. In a metal there are already so many electrons that it is almost impossible to change the numbers.

Although the vacuum tube continues to play an important role in electronics, the transistor has several inherent advantages. It operates without any warm-up time and requires merely a fraction of the power that a vacuum tube needs. It is extremely reliable: when properly manufactured and operated within its design limits, the transistor can operate trouble-free for decades. It can also transmit very high frequency signals at very high speeds, making it suitable for use in modern switching and computer systems, and can be made smaller than a pinhead, allowing extreme miniaturization.

Semiconductors, neither good insulators nor good conductors, first gained notice because of their photoelectric properties and their use in rectifiers where they conduct more easily in one direction than the other.[1] As early as

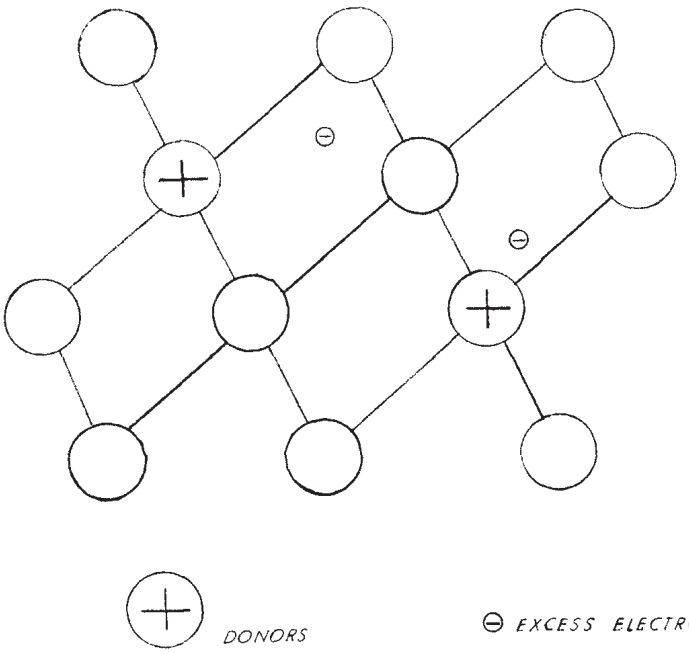


Fig. 1. *n* type semiconductor.

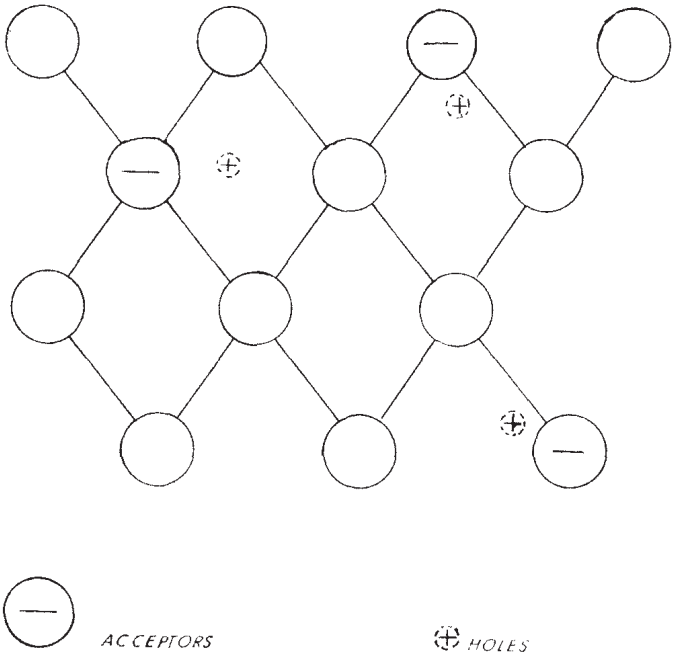


Fig. 2. *p* type semiconductor.

1920, serious attempts to understand the electrical properties of these solids were begun. Many scientists all over the world were interested in the problem. The advent of quantum mechanics gave us the first real understanding of how atoms stick together in solid crystals. Quantum theory also gave us understanding of how electrons in metals were free to move and conduct electricity, and A. H. Wilson showed us how this theory could explain insulators and semiconductors. Valence electrons in insulators and semiconductors are not free to move under the influence of an electric field. However, if a few valence electrons are excited thermally or by light, they can move; and the vacant holes in the bonding structure can also move and act as positively charged carriers of electric current.

In 1931, my supervisor and colleague at Bell Telephone Laboratories, J. A. Becker, and I first started work on copper oxide, one of the then better known semiconductors. (The Bell System was interested in this material because of its rectifying properties and possible use as a modulator of electrical communications signals.) There were two main problems—flow of current inside the solid and flow of current at contacts between the semiconductors and metals. At such contacts current would flow more easily in one direction than the other (retification), and also at such contacts illumination of the surface would sometimes produce a voltage (photo emf). A light shown on the main body of the semiconductor would also increase its electrical conductance (photoconductivity).

Early in 1940, Mervin J. Kelly, who was then Director of Research for Bell Telephone Laboratories, called us into his office to witness a demonstration by Russell S. Ohl, a member of the staff who was working with silicon, a then little known semiconductor. Ohl showed us a small rectangular piece of a black solid with two metal contacts. When light from a flashlight illuminated a narrow region near the middle of this piece of silicon, a photoelectromotive force (emf) of about 0.5 V was developed. This was hard to believe. In the first place, the contacts were not being illuminated and the photo emf was ten or more times larger than any we had ever seen. Moreover, the silicon was black; that is, opaque to visible light. In fact, I did not believe what I saw until Ohl gave me a piece to work with in my own laboratory.

This was the first *p-n* junction. Some time before this, G. C. Southworth had been having trouble getting something to detect the short radio waves (microwaves) with which he was working. Vacuum tubes did not work well at these wavelengths. He remembered the old cat's whisker detectors that were used before vacuum tubes were invented and decided to give them a try. He went to the second-hand radio market in the Cortland Street section of New York and there found some cat's whisker detectors, dusted them off, and tried them. They worked.

Hearing about it, Russell S. Ohl, of Bell Telephone Laboratories, became interested in the phenomenon. These detectors were semiconductors, made of either silicon or galena (lead sulfide). Ohl decided to concentrate on silicon. The silicon you could buy was very nonuniform in its detection properties. This detection involved rectification of the radio wave at the point of contact of the metal cat's whisker with the silicon. Sometimes the rectification would be in one direction and sometimes in the opposite direction, sometimes not at all. Ohl asked the metallurgists at Bell Labs to see if they could not make the silicon more uniform by purifying it. J. H. Scaff and H. C. Theuerer soon found they could purify the silicon by melting it in high vacuum. Sometimes the ingots that they made would rectify all one way and another ingot would rectify all the other way. Those that conducted best when the silicon was negative were called *n* type (Fig. 1) and the other *p* type (Fig. 2). One ingot that

Scaff and Theuerer gave Ohl was all *n* type on one end and all *p* type on the other end, and a piece cut out to include the boundary between these two regions was the one Ohl had demonstrated.*

Scaff and Theuerer went on to find that the conductivity in silicon was mainly due to very small traces of impurities. These impurities were elements in the fifth column of the periodic table, like phosphorus, which gave excess electrons when added and made silicon *n* type. Elements in the third column such as indium gave excess holes and made silicon *p* type.**

With the advent of World War II it became important to make good silicon detectors for radar. This work was carried forward by Scaff, Theuerer, Ohl, and by many others outside the Bell Laboratories. Much independent work was done in England and at the Radiation Laboratory at MIT. Lark-Horovitz and his group at Purdue worked with germanium, another semiconductor like silicon (both are in the fourth column of the periodic table). Others, including myself, worked on entirely different war problems.

Just before the War, N. F. Mott in England and W. Schottky in Germany independently developed the first good theory of rectification; B. Davydov in the USSR also contributed. When two electrically conducting solids are in contact and in equilibrium there is an electrostatic difference in potential between the two solids. This means that there must be a charge double layer at the contact. Since the free charge density, electrons or holes, in the semiconductor is very small this charge layer must penetrate into the semiconductor. This theory suggested that if the potential difference is the right sign there will be a potential hill at the surface of the semiconductor for, say, electrons and if the semiconductor is *n* type the contact will consequently rectify. If it was *p* type it would not rectify, and one would have to use a metal giving a potential difference of the opposite sign to get retification.

Another point of interest was that this rectification at the metal semiconductor contact was analogous to the recification obtained in a vacuum tube diode; many people thought of putting in a third element like a grid in the vacuum tube triode, to make an amplifier, but none of the schemes worked.***

* When *n* and *p* type silicon are in contact as in Ohl's junction the holes tend to diffuse over to the *n* side charging it positively, and electrons diffuse the other way charging the *p* side negatively. This results in an equilibrium electrostatic potential preventing any further diffusion. If the density of holes and electrons were both equal on the two sides of the junction there would be no diffusion; therefore, no charging and no potential. Shining light on the junction creates equal numbers of extra electrons and holes, thus tending to reduce the tendency to diffuse and thus reduce the electrostatic potential. This change in potential from the equilibrium value is the photo emf. W. H. Brattain, *Science,* **126**, 151 (1957).

** These impurities were below the level of spectroscopic detection at that time. The odor of the ingots as they came out of the oven was traced to phosphorus by Scaff and Theuerer.

*** R. Hilsch and R. W. Pohl succeeded in putting a grid in an alkali halide crystal but the crystal would only respond to signals longer than 1 sec.

In January 1946, because of Bell Labs' interest in the use of semiconductors in circuit devices and the possibility of a solid state amplifier, scientific research to enable us to understand semiconductors was resumed in earnest. The research team was headed by William Shockley and S. O. Morgan. Shockley, a theoretical physicist, worked as part of the research team he headed. John Bardeen, also a theoretical physicist, joined the group (Fig. 3).

At the first meeting of the group it was realized that in spite of all the work done before and during the war, we were still far from a real understanding. One reason was that copper oxide and other semiconductors on which the early work had been done were very complicated solids. Silicon and germanium were the simplest, and the decision was to try to understand these first. Our work was directed toward a fundamental understanding of the problem though we were well aware of the technical importance of a semiconductor amplifier if one could be made.

Besides those already mentioned there was G. L. Pearson, who was primarily interested in the bulk properties of semiconductors, R. B. Gibney, a physical chemist, and H. R. Moore, a circuit expert who aided greatly in making measurements and devising novel circuits for our experiments. I was primarily interested in what went on at the surface where contact was made or at the boundary between *n* and *p* types. Of this group only Pearson and I had worked extensively with semiconductors before the war.

Scaff and Theuerer were in a position to furnish us with polycrystalline ingots of either *n* or *p* type silicon or germanium of any specified resistance. This, of course, was a great help. Based on the Mott-Schottky theory of rectification, Shockley had come to the conclusion during the War that it should be possible to control the density of electrons near the surface of the semiconductor by means of an electric field applied between the surface and a

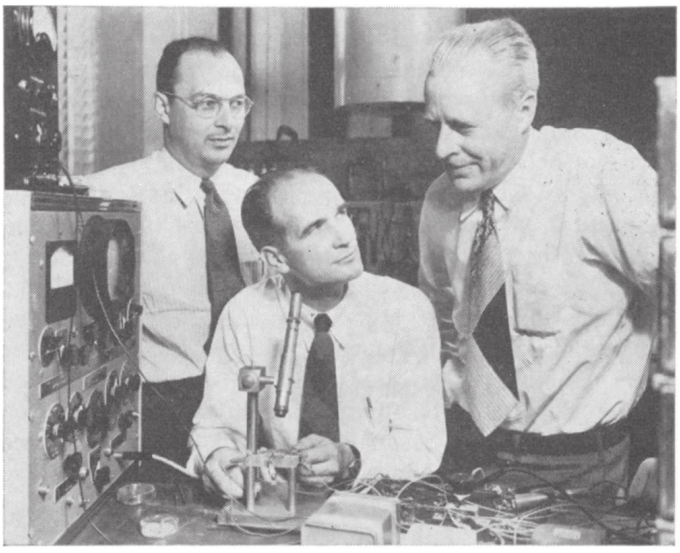


Fig. 3. Nobel Prize winners John Bardeen (left), William Shockley, and Walter H. Brattain at Bell Telephone Laboratories in 1948 with apparatus used in the first investigations that led to the invention of the transistor. The trio received the 1956 Nobel Physics Award for their invention of the transistor.

metal electrode insulated from the surface. If this were true one could vary the conductivity in the surface of a thin wafer semiconductor by means of the field and thus make an amplifier. Many experiments were devised by Shockley to test this hypothesis, but the effect was several orders of magnitude less than predicted.

Moreover, several predictions based on the theory did not agree with experimental results. In order for two electrical conductors to be in equilibrium there must, in general, be an electrostatic potential difference between them (contact potential). The prediction was that there should be such a potential between *n* and *p* type silicon and between *n* and *p* type germanium. Experiments showed that this potential was very small, almost zero. Different metals should have different contact potentials with respect to the semiconductor. If the sign of this potential was not right there should be no rectification at the point of contact; yet, experimentally, all metal points worked more or less equally well. Additionally, S. Benzer at Purdue found that contact between two pieces of germanium, one *n* type and one *p* type, did not act as one would expect. I believe the group as a whole slowly realized that these results were all of a piece, and it was Bardeen who successfully explained them all by applying to this problem the concept of surface states; that is, that the electrons could be trapped at the semiconductor surface, and that the semiconductor was in equilibrium with its surface before any electrical contact was made to it. This, of course, implied a space charge layer in the surface of the semiconductor equal and opposite to the charge trapped on the surface. Consequently, the electrostatic potential change between the interior of the semiconductor and the surface which was necessary for rectification was a property of the semiconductor and its surface—independent of the metal contact. This theory immediately suggested new experiments.

When any novel ideas or new experimental results were obtained by any member or part of the group it was our practice to call the group together for a presentation and discussion. These sessions were inspirational, and it was during such sessions that many of the better suggestions for further work were made. Two suggestions for experiments came about as a result of the meeting and discussion that occurred on the occasion of Bardeen's presentation of his theory. Shockley suggested that, if trapping centers for electrons on the surface were limited in number, one should be able to measure some small change in contact potential between *n* and *p* type samples of, say, silicon as the samples became more strongly *n* type and *p* type through the introduction of more and more of the proper impurities. I suggested that, if extra electrons and holes were excited by illumination in the surface region (where there must be an electric field due to the space charge), the electric field would tend to separate the excited electrons and holes, thus changing the surface charge and contact potential. Both these experiments were successfully performed.

Another suggestion was to try to reduce the temperature low enough so electrons trapped at the surface could be frozen and a field effect observed. Experiments by Pearson and Bardeen showed that this was the case.

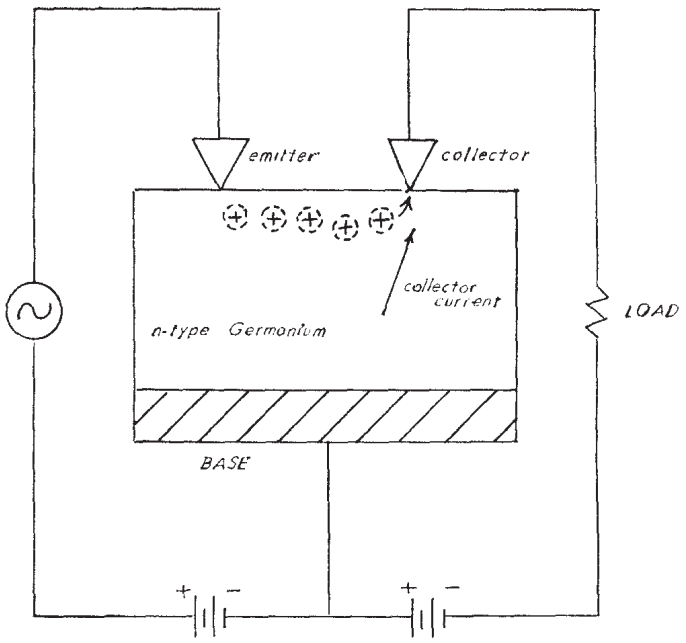


Fig. 4. Point contact transistor at work.

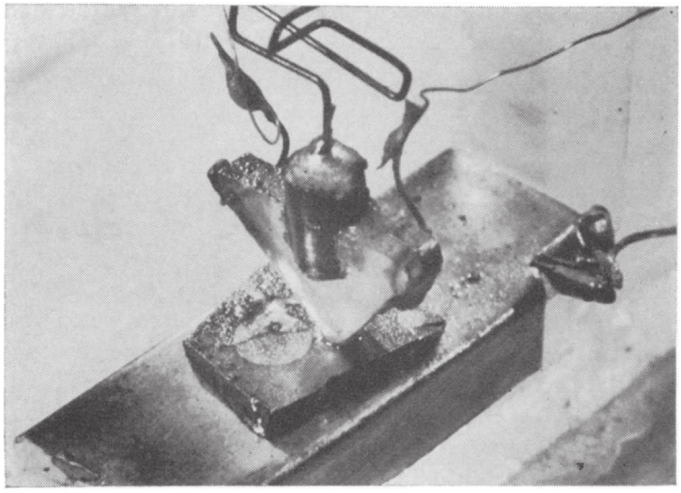


Fig. 5. The original transistor.

Another experiment that I tried was to measure the change of potential at the germanium or silicon surface as a function of temperature. Condensation of moisture from the air on the cold semiconductor surface interfered with this experiment. As a result it was decided to try immersing the system in an insulating liquid. The apparatus had been arranged to measure contact potential and photo emf's, and when the liquids were tried large changes in photo emf's were observed. Some of the liquids tried (such as water) were not strictly insulating but were electrolytes. When I was showing these phenomena to Gibney he suggested varying the potential between the semiconductor surface and the reference electrode. When using an electrolyte we could make the photo emf very large by this means. By changing the sign of the potential we could make the photo emf go through zero and change sign. It was recognized that this was, in essence, Shockley's field effect. By using the electrolyte we could vary the space charge layer and potential inside the semiconductors at the surface.

These results were presented to the group as a whole and one morning one or two days later Bardeen came into my office with a suggested geometrical arrangement to use this effect to make an amplifier. I said let's go out in the laboratory and do it. We covered a metal point with a thin layer of wax, pushed it down on a piece of *p* type silicon that had been treated to give an *n* type surface. We then surrounded the point with a drop of water and made contact to it. The point was insulated from the water by the wax layer. We found as expected that potentials applied between the water and the silicon would change the current flowing from the silicon to the point. Power amplification was obtained that day!

Bardeen suggested trying this on *n* type germanium, and it worked even better. However, the water drop would evaporate almost as soon as things were working well, so at Gibney's suggestion we changed to glycol borate, which hardly evaporates at all. Another problem was that amplification could be obtained only at or below about 8 Hz. We reasoned that this was due to the slow action of the electrolyte. Optimum results were obtained with a dc negative bias on the electrolyte when using *n* type germanium. Under these conditions we noticed an anodic oxide film being formed under the electrolyte. We decided to evaporate a spot of gold on such a film and, using the film to insulate the gold from the germanium, use the gold as a field electrode to eliminate the electrolyte. The film was formed, the glycol borate washed off, and the gold spot with a hole in the middle for the point was evaporated. When this was tried, an electrical discharge between the point and the gold spoiled the spot in the middle; but by placing the point around the edge of the gold spot, a new effect was observed. In washing off the glycol borate, we had inadvertently washed off the oxide film which was soluble in water. The gold had been evaporated on a freshly anodized germanium surface. When a small positive potential was applied to the gold, holes flowed into the germanium surface greatly increasing the flow of current from the germanium to the point negatively biased at a large potential! Four days later, on 23 December 1947, two gold contacts less than two thousandths of an inch apart were made to the same piece of germanium; and the first transistor was made (Figs. 4 and 5). This was an eventful day, and here is how I recorded it in my notes:

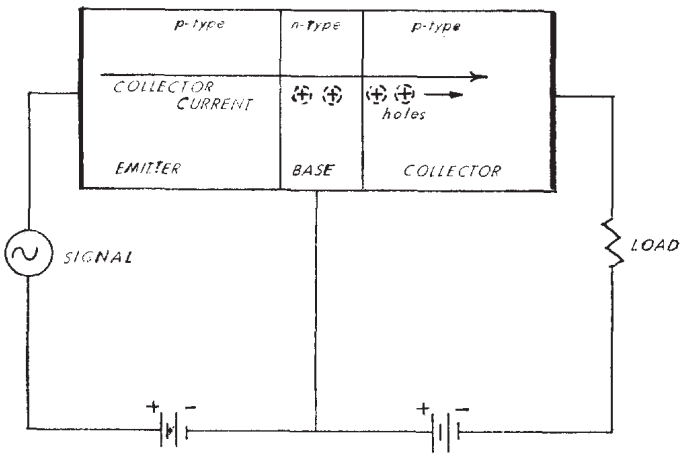


Fig. 6. Junction transistor at work.

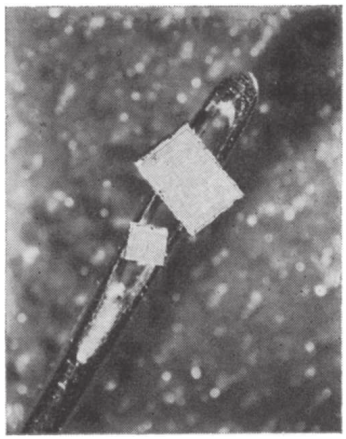


Fig. 7. Microminiaturized modern transistors enlarged 200 times.

"Using the germanium surface (see top of page 197) and the gold contacts the following circuit was set up. This circuit was actually spoken over and by switching the device in and out a distinct gain in speech level could be heard and seen on the scope presentation with no noticeable change in quality. Various people were present and witnessed this test and listened—of whom some were the following: R. B. Gibney, H. R. Moore, J. Bardeen, G. L. Pearson, W. Shockley, H. Fletcher and R. Bown. H. R. Moore assisted in setting up the circuit and the demonstration occurred on the afternoon of 23 December 1947."

The transistor was named in my office by J. R. Pierce. Bardeen and I wanted a name that would fit in with varistor and thermistor, but we were more or less devoid of ideas when I presented the problem to Pierce. Pierce knew that the point contact transistor was the dual of a vacuum tube, circuit wise. After some thought Pierce mentioned the important parameter of a vacuum tube—transconductance—then a moment later its electrical dual —transresistance. Then he said "transistor" and I said, "Pierce, that is it."

The original transistor was called a point-contact transistor because it was essentially a wafer of germanium with two pointed wire contacts made close together on one side. Shockley went on to work out the theory of the *p-n* junction and to predict the *p-n-p* or *n-p-n* transistor (Fig. 6).

When in 1950 G. K. Teal and J. B. Little successfully made single crystals of germanium, it soon became possible for Morgan Sparks to make a junction transistor at Bell Labs. A few years after the alloy transistor was developed, and later the diffused base transistor made it possible for a single wafer of semiconductor material to contain thousands of transistors.

The same knowledge and theories that led to the transistor's development have led to the development of whole new families of solid state devices. This in turn has brought about a remarkable technology called microelectronics. This new technology has shrunk transistors and other circuit elements to dimensions almost invisible

to the unaided eye. For example, a complete microcircuit consisting of ten to twenty transistors and forty to sixty resistors can be interconnected and built onto a piece of silicon only about a tenth or twentieth of a square inch (Fig. 8).

Today, then, the transistor is no longer a single device. It is a family of devices made in dozens of different ways and used in almost every kind of electrical circuit. When the work that led to the transistor effect was being done, however, my chief interest was in understanding solid state surfaces. The transistor actually came into being because of our work in solid state physics, specifically in the field of semiconductor physics.

The transistor came about because Bell Laboratories undertook to develop fundamental knowledge to a stage where human minds could understand phenomena that had been observed for a long time.

I believe the discovery of the transistor effect and the invention of the transistor illustrate the power of scientific research in the areas of science relevant to a purpose. My personal satisfaction came mainly from the basic discoveries we made, but I must also admit no little pleasure in the fact that my work also has had widespread practical application for the benefit of man.

Reference

1. References to many of the events mentioned in this article can be found in G. L. Pearson and W. H. Brattain, Proc. Inst. Radio Engrs. 43, 1794 (1955), where the early history is told in greater detail.